Chapter 2: Data Storage in Enstore

2.1 Storage Groups

Each experiment or research project is assigned a unique *storage group* identifier by the Enstore administrators. Enstore uses the storage group names to control and balance assignment of resources, such as tape drives and media, among the experiments. Each storage group is assigned an area in PNFS, e.g., an experiment XYZ might be assigned the storage area `/pnfs/xyz`.

2.2 File Organization on Storage Media

2.2.1 File Family

Files are grouped on data storage volumes according to a *file family*[1] attribute. A file family is a name that defines a category, or family, of data files. Each experiment (i.e., each storage group) must carefully plan its set of file families. There may be many file families configured; by design there is no pre-set upper limit on the number. A given storage volume may only contain files belonging to one file family.

Every directory in `/pnfs` namespace has a file family associated with it. Every data file added to the Enstore system (i.e., every file for which an entry appears in the `/pnfs` namespace) is thus associated with the file family of the directory under `/pnfs` into which it was initially copied.

Associated with a file family are a *file family width* (an integer value), and a *file family wrapper* (a format specification). The file family, file family width, and file family wrapper for a PNFS directory are initially inherited from the parent directory. They may be reset as permissions allow, generally only by a small group of designated people in each experiment.

[1] The grouping is really based on the triplet of quantities: storage group + file family + wrapper, collectively called a "volume family". However, most users have access to only one storage group, and use the default wrapper, so from the user's perspective, the only relevant attribute for file grouping is file family, typically.

2.2.2 File Family Width

File family width is an integer value associated with a file family that is used to limit write-accessiblity on data storage volumes. There is currently no width associated with reading. For a given media type and for a given file family, Enstore limits the number of volumes available for writing at any given time to the value of the file family width (except when unfilled volumes are already mounted for previous reads). Correspondingly, the number of media drives on which the volumes are loaded is also limited to the width.

2.2.3 File Family Wrapper

A file family wrapper specifies the format of files on the storage volume. It defines information that gets added before and after data files as they're written to media. In this way the data written to tape is self-contained and independent of metadata stored externally.

There are three wrapper types implemented, `cpio_odc`, `cern` and `null`. Currently (September 2005) most tapes are written using the `cpio_odc` wrapper. The `cpio_odc` wrapper is the default wrapper set up by the Enstore admin when a new PNFS area is created.

- All files with the `cpio_odc` wrapper are dumpable via **`cpio`**. This wrapper has a file length limit of (8G – 1) bytes. It is sufficient for the vast majority of data files, as most files are still under 2GB.
- The `cern` wrapper accommodates data files up to (10^21 –1) bytes, which in effect limits the filesize to the tape length, since spanning and striping are not supported as explained in section 2.4 *Data Storage Volumes in Enstore*. It matches an extension to the ANSI standard, as proposed by CERN, and allows data files written at Fermilab to be readable by CERN, and vice-versa[1].
- For NULL volumes there is a `null` wrapper. See section 6.5.8 *NULL File Directories* for other restrictions on using NULL volumes.[2]

[1] Since CERN will allow files to span volumes, and Fermilab doesn't, users will not be able to use Enstore to read volumes from the CERN system that contain partial files.

[2] On rare occasions the Enstore administrator may determine that the null wrapper should be used for tapes written elsewhere and imported into Enstore.

2.3 File Size Limitations

Enstore limits the size of a data file to the tape capacity, i.e., a single file cannot span more than one volume. The `cpio_odc` wrapper further limits the file size to (8G – 1) bytes, as mentioned in section 2.2.3 *File Family Wrapper*. Your OS may restrict your files to yet a smaller size.

Tape sizes are as follows (all sizes shown are for non-compressed tapes):

LTO-2	200 GB
LTO-1	100 GB
9940B	200 GB
9940	60 GB
9840	20 GB (no longer used)
DECDLT (4000)	20 GB Compressed (Special read-only situation)
Mammoth 2	60 GB (no longer used)
Mammoth	20 GB (no longer used)

Wrapper size limitations are as follows:

cpio_odc	8 GB -1B; (about $8*10^9$ Bytes)
cern	About 10^{12} GB; or 86.7 Exabytes; (10^{21}-1 Bytes)
null	N/A; not available for writing to tapes; used for special situations

☞ Note: PNFS can only represent a data file’s size accurately up to (2G–1)B; beyond that, the file size is shown as 1. Enstore knows, stores and uses the real file size, so this PNFS display limitation does not pose a functionality problem. A 1 byte file size indicated by PNFS indicates to the the user that the file is likely quite large. (You can use the **`enstore pnfs`** command with the **`--filesize`** option as described under section 9.4 *enstore pnfs* to find the actual file size.)

2.4 Data Storage Volumes in Enstore

Enstore is designed to support a variety of data storage media. Currently, only tapes have been implemented, and the information in this section has been written with tapes in mind. In principle, the information should apply equally to other media types; we will update this section as necessary when other media types are implemented.

2.4.1 Tape Features

Tapes are self-describing and exportable so that in the unlikely event of lost metadata in Enstore, a volume can be dumped, and the information retrieved.

Tapes are required to have ANSI Vol1 header labels. Labelling helps to easily identify each volume and/or test for a blank one, thereby inhibiting the inadvertent overwriting of used tapes. The Enstore admin needs to know whether tapes are labelled or completely blank when they are inserted into the library; the Enstore software can label tapes if necessary.

2.4.2 File Organization, Storage and Access

Data files are physically clustered on volumes according to each experiment's file family classification scheme. A given volume may only contain files belonging to one file family, one wrapper type, and one storage group.

A single file cannot span more than one volume[1]. When writing files to media, Enstore compares the size of the file it's ready to copy against the volume's remaining empty space in order to determine whether the file will fit. If the file is too large to fit, Enstore marks the volume as full, and writes the file to a different volume. Thus volumes are filled, modulo some fraction of a data file. (Volumes can be reopened if an administrator decides that too much space is left unused.)

Enstore supports random access of files on storage volumes, and also *streaming*, the sequential access of adjacent files.

[1] Since files cannot span multiple volumes, striping is not supported either. Striping refers to files (usually large ones) being split onto two or more volumes, each writing simultaneously, in order to expedite the writing process.

2.4.3 Quantity of Volumes

Enstore allows data storage volumes to be “faulted out to shelf”, i.e., removed from a robot. This feature makes it possible for each experiment to have a larger number of volumes than it has slots in the robot, and in fact there is no limit on the number of volumes used by an experiment. To accommodate the unmatched numbers of volumes and slots, Enstore provides separate quotas in volumes and in slots.

Note that moving volumes in and out of the robot requires operator intervention, and should be minimized.

2.4.4 Import/Export of Volumes

Tapes can be generated outside the Enstore system and imported into it. Conversely, Enstore tapes can be dumped via standard UNIX utilities, thereby allowing them to be readable with simple tools outside the Enstore framework. The tools to do this are wrapper-dependent (e.g., tapes whose files have the cpio wrapper can be dumped via the **`cpio`** utility). Currently there is no utility (except **`dd`**) to dump a tape whose file family wrapper is cern.